How to reach top accuracy for a visual pedestrian warning system from a car?

Floris De Smedt, Steven Puttemans and Toon Goedemé
KULeuven, EAVISE
e-mail: floris.desmedt@kuleuven.be, steven.puttemans@kuleuven.be, toon.goedeme@kuleuven.be

***Abstract*— Due to the wide applicability of pedestrian detection in surveillance and safety, this research topic has received much attention in computer vision literature. However, the focus of this research mainly lies in detecting and locating pedestrians individually as accurate as possible. In recent years, a number of datasets are captured using a forward looking camera from a car, which imposes the application of warning the driver when pedestrians are in front of the car. For such applications, it is not required to detect each pedestrian independently, but to generate an alarm when necessary. In this paper we explore techniques to boost the accuracy of recent channel-based algorithms in this application: algorithmic refinements as well as the inclusion of an LWIR image channel. We use the KAIST dataset which is constructed from image-pairs of both the visual and the LWIR spectrum, in day and night conditions. We study the influence of techniques that have shown success in literature.**

***Keywords*— Pedestrian Detection, LWIR, ACF, System Test**

I. Introduction

Due to the wide applicability of pedestrian detection in a variety of applications, including traffic, surveillance, robotics and safety, there has been a lot of research attention on this topic. In the last decade, we could observe a drastic improvement from a miss rate of 68%, using the HOG detector [1], to under 10% miss rate [2] on the challenging Caltech dataset [3]. These recent results could be encountered by extending a strong ACF baseline with the use of convolution masks, applied on the feature information, and the additional use of Convolutional Neural Networks (CNNs).

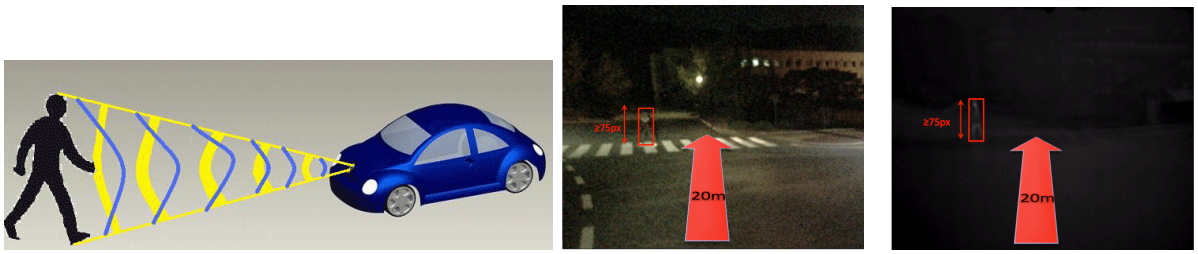
In this paper we use these recent pedestrian detection algorithms as main ingredient of a visual pedestrian collision warning system for a car driver. Our system test validates that an alarm will be generated if pedestrians are too close in front of the car (as illustrated in figure 1). For our experiments we make use of the KAIST dataset [4], which contains images for both day and night conditions. We clearly see that color based detectors can not handle the night conditions properly, and that the additional use of LWIR can greatly compensate for this lack of detection quality. We show that with a few modifications to the training process, an ACF model can be trained that outperforms the current state-of-the-art on this dataset, while still maintaining high detection speed.

In section II we give an overview of the related work on the topic of pedestrian detection. In section III we describe how we elevate this to a complete pedestrian collision warning system, which we use to compare the influence of different training choices in section VI. We extend the default ACF detector by incorporating features from the LWIR spectrum in section IV. In section V we study different training parameter choices that have shown beneficial effect in literature for detectors evaluated on the Caltech dataset, to study its influence on the detection quality of our detector on the KAIST dataset. In section VI we demonstrate the influence of these choices have on both the reliability of the warning system and the detection speed. Finally we conclude in section VII.

II. Related Work

In 2005, Dalal and Triggs proposed the use of Histogram of Oriented Gradients (HOG) for the use of human detection [1]. The impressive results they obtained at the time showed the potential of HOGs as a feature, and even today it is used in state-of-the-art pedestrian detection techniques. To improve the accuracy further, we can recognise two approaches. In 2008 Felzenszwalb et al. [5] proposed the use of a deformable model (DPM), where next to a model for the object as a whole, also part models where used, to detect also separate parts of the object (e.g. the limbs). By allowing a small deviation of the positions of the part models relative to the root-model, the model is *deformable*. The accuracy gain obtained with the allowed flexibility of the object came with a large computation cost however. In 2010 Felzenszwalb et al.[6] proposed the use of a cascaded model, which partly solved this issue with a 10x speed-up on average over a broad range of object categories. Although the use of deformable models is used in detectors reaching high accuracy results [7], this allowed flexibility is not a necessity to reach top performance on pedestrian detection [8].

In contrast to using a more complex model-structure, the type of features could be extended, as performed by Dollár et al. [9] in their Integral Channel Features (ICF) detector. Here the use of HOG features is extended with the use of color features, forming HOG+LUV. In contrast with the framework of Viola and Jones for face detection [10], on which the ICF detector was based, they only considered the use of randomly selected first order features at the time. In 2013 the training process of the ICF detector was reevaluated, leading to a large accuracy improvement [11], forming the *Roerei* detector. A large part of this improvement was obtained by considering all possible rectangular-shaped features inside the model window, instead of only a pool of randomly selected features as used by ICF. The required computational power for training however made this approach infeasible to use in practice (several days



(a) Typical system setup (b) (c)

Fig. 1: System test applied on an image-pair (color + LWIR) of the KAIST dataset.

for a single model on a high-end computing server). A reduction to only using all possible squares instead, still obtained high accuracy at an acceptable computational requirement. This extention of the ICF detector was coined *SqrtChnFtrs* (Squared Channel Features) and formed the baseline detection approach for the *Katamari*-detector [8]. In 2014, Dollár et al. [12] published a generalization of their "Fastest Pedestrian Detector in the West" (FPDW) [13] approach to improve the computation time of the feature pyramid. Instead of calculating each layer from raw image data, they proposed the use of feature approximation instead. By combining this approximation approach with a use of a fixed size of feature rectangles (a reduction of the feature pool used by SqrtChnFtrs), led to a high speed and accurate pedestrian detector, which was coined the Aggregate Channel Feature detector (ACF). The training process of this detector is performed by using a highly optimised AdaBoost implementation [14]. Among other optimisations, this paper states that the same accuracy, or at least a very close approximation, can be obtained by using only a fraction (6.25%) of the feature pool for training.

Wan et al. [15] further improved accuracy by convolving each of the 10 used LUV+HOG channels (6 gradient orientations, 3 color LUV channels and the gradient magnitude) with 4 convolution masks, based on a PCA analysis of these features, leading to 40 channels. This approach was in 2015 extended by Zhang et al. [16], by replacing the LDCF convolution masks with 39 checker-board patterns, obtaining the most accurate detection approach so far, without the use of Deep Learning, shown on the Caltech dataset. For the sake of reducing training time, recently they proposed the use of only 9 thoughtfully chosen convolution filters [2], with only a single percent miss-rate drop in accuracy compared to the 39 filters of the *Checkboard* detector. This small drop in accuracy was thoroughly compensated for by using improved annotations and an additional classification step of the detections, using a CNN network using the VGG CNN architecture [17].

In 2015, Hwang et al. [4] pointed out the lack of detection quality of color based pedestrian detection techniques on night images. To solve this, they proposed the use of Long-wave infrared (LWIR) next to the use of traditionally used color images. To evaluate this approach, they captured a very large dataset, coined the KAIST dataset, containing a calibrated image pair of color and LWIR (figure 1) for both during day and night time. They made use of the ACF detector, which they extended with additional channels calculated on the LWIR image input. In this paper we improve further on this work by creating a stronger baseline ACF implementation, paving the way for a similar detection quality evolution that has taken place in pedestrian detection in the visual spectrum.

Using the sliding-window paradigm, the previously discussed detectors perform an evaluation of the detection model on each location of the image, and this at multiple resolutions of the source image to cover multiple sizes a pedestrian can appear at. It is however possible to reduce this search space with the use of scene constraints. The most applicable scene constraint for the conditions we will work with, is the use of a ground constraint. This constraint implies that a certain height of pedestrian can only appear in a part of the image. This was demonstrated by [18] where the homography of the ground plane was exploited for this purpose. In [19] however, De Smedt et al. showed that this ground constraint relation can be approximated by a first order linear function modelled on a training set, which they demonstrated on the Caltech dataset.

III. System test construction

Using a system test, we determine the accuracy when a pedestrian detector is used as a warning system in a driving car. Instead of determine if each pedestrian is accurately localised on a frame-by-frame basis, we validate if an alarm will be generated when pedestrians walk in front of the car. When a car is driving at 50km/h, roughly 12.5m to 18.75m is required to stop the car, according to the rule of thumb given in equation 1 and 2, which give the breaking distance *BD* in metres in function of a speed *v* in km/h.

$$BD_{dry} = \frac{(\frac{v}{10})^2}{2} \tag{1}$$

$$BD_{wet} = BD_{dry} \times 1.5 \tag{2}$$

According to [4], a pedestrian at the size of 75px is approximately 20m in front of the car. To brake on time, the processing has to be performed between 1.85fps and 11.11fps, for respectively dry and wet weather conditions, when driving the aforementioned 50km/h. Note that for a warning system, the level of occlusion is not important, so for our experiments we have to generate an alarm independent of the occlusion

level of the annotations (where traditionally the pedestrians have to be at least 65% visible).

For our algorithmic comparison in section VI we use an ROC-curve, which visualises the True Positive Rate (TPR), as given by equation 3, in function of the False Positive Rate (FPR), as given by equation 4.

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

Hoedemaeker et al. [20] state that a FPR of 5% is acceptable for such a safety system to support the driver. Based on this, we will compare algorithms using the TPR they reach at a FPR of 5%. Our evaluation on this dataset differs from the default, since we will use each 5th image instead of each 30th, to guarantee a good sampling.

IV. Exploiting the LWIR-spectrum

Figure 2 shows that the performance of the ACF algorithm when only images from the color spectrum are taken into account from both day and night conditions of the KAIST dataset [4]. As we will see in figure 10 is this performance far from sufficient for night conditions. Less than a quarter of the alarms is generated when necessary, having a false alarm rate of 5%.

Information from the LWIR-spectrum can largely benefit the detection accuracy during night conditions. To use the LWIR-spectrum, a sensor is used that registers wavelengths between 8μm and 14μm. This range of the spectrum allows to register heat radiation, which is less sensitive to the presence of fog, dust and night conditions, compared to the visual spectrum. Currently, the price of these sensors is drastically reducing, which allows the use of LWIR sensors also in low-cost applications. To extend the default ACF-detector with information of the LWIR-spectrum, we use a very similar approach as [4]. The HOG+LUV channels, calculated on the color image, are concatenated with the same type of channels (raw pixel intensity information and gradient information) but calculated on the LWIR image, forming a total of 18 channels (10 from the color spectrum as used before + 6 LWIR gradient orientation + 1 LWIR gradient magnitude + 1 LWIR intensity) Since the algorithm itself is not altered, the many extensions that exist for channel-based detectors for both speed [21] and accuracy [15], [16], [2] can be applied on this combined detector also.

Figure 2 compares the detection results of the color-based ACF model (*ACF-Color*) with an ACF model incorporating the LWIR channels (*ACF-Both*) and the best results obtained by [4] (*ACF-T-THOG*). Note that our combined detector, which uses the same code as used in the color information to calculate the additional 8 channels on LWIR, already outperforms the combination of [4] (*ACF+T+THOG*) with a large margin (almost 8% miss-rate). As we will see in section VI this improvement in detection quality is also beneficial for the results of the system test.

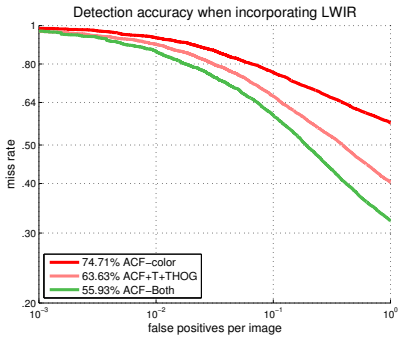


Fig. 2: Performance when using only channels from the visual spectrum in comparison with the current state-of-the-art.

V. ACF model improvements

In section II we already discussed a range of improvement on detection quality for channel-based detectors, in most cases validated on the Caltech dataset. In this section we study if these techniques will have the same benefit on the KAIST dataset as an extension of our *ACF-Both* detector. In section VI we study the influence of these improvements on the results of the system test.

A. Using ACF+

A first extention of our *ACF-Both* detector we propose, is the use of an ACF+ [15] model training setting instead of ACF. Table 1 gives an overview of the parameters used for training both models. Since ACF+ uses more negatives, stronger weak classifiers and allows more weak classifiers to be used in the model, it is capable of modelling a more complex decision surface for classification. In figure 3 we see that using the ACF+ settings leads to a large accuracy gain over the classic ACF setting, with a decrease of 7.5% in miss-rate.

Table 1: Comparison of training parameters of ACF and ACF+

Parameter	ACF	ACF+
# Weak classifiers	2048	4096
Max tree depth	2	5
# negatives	5 000	25 000
# accumulated negatives	10 000	50 000

B. Influence of the training set

During the training process, a decision surface between positives (pedestrians) and negatives (background) is searched for based on example images. This means that when using a similar dataset during training as will be used during testing time, will in most cases lead to the highest accuracy [8].

In this subsection we study if an accuracy gain can be obtained by using a separate model for only day and only

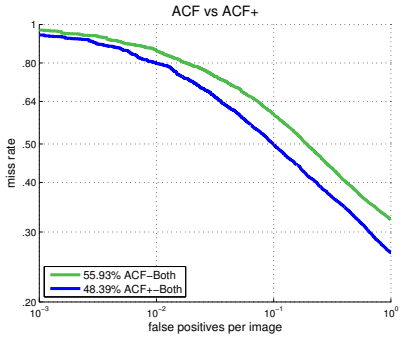


Fig. 3: Comparison of ACF and ACF+

night images, instead of a single model trained on both, as was used earlier. In figure 4 we compare the detection accuracy of ACF+ models trained on only day images, only night images and the combination of both (as shown in the legend of these figures). As we can observe, is the ACF+ model setting we use sufficiently flexible to create a model that works in both day and night conditions. It is remarkable however that this combination does even perform better than a model that is trained for the specific setting (day or night). This advantage could both be caused by the larger amount of training data, which helps to better generalise the data, and/or that the day and night images have complementary information that turns out to be beneficial during testing time.

In figure 5 we show the feature selection over the different feature types. In this visualisation, we normalised each feature value both based on the score contribution to the decision tree the feature is part of, and the decrease in classification error it implies on the training set. Note that in the three models, the raw pixel data of the color image is more used than those of the LWIR image. As can be expected are the color based features more used when a model is only trained on day conditions, compared to when night time images are included in the training set.

We can conclude that the use of separate models for day and night will not lead to an accuracy improvement, so training a model on both the day and night images is still the best choice.

C. Amount of training data

It is shown earlier in literature that using more training data can be beneficial to reach higher accuracy [8], [16], [2]. Two conditions have to be fulfilled however 1) the model should be capable of capturing this information, and 2) The additional positive examples should contain information that is not already available. This first condition is the strenght of Deep Learning Models [22], [2] where a lot of data is used to train these very strong network architectures. By using the ACF+ model setting, we have increased the strength of the model in the spirit of the first condition. The second condition can better be explained by representing the trained model as a decision surface, and additional examples are only beneficial if these are close to this decision surface. When using SVM machine learning, only the examples close to this decision surface are taken into consideration during training [23] (coined the Support Vectors). In practice the useful examples will be those that are classified incorrectly (FN and FP), who are positioned at the wrong side of the decision surface, since these are the ones that contain information not already captured by the model.

In figure 6 the resulting accuracy is shown when using each 4th image for training instead of each 20th, of which the latter is the default training setting for this dataset. Also here, however, we have to conclude that using more training data does not further improve the accuracy of our ACF+ model, implying that the training information of the additional frames is redundant, or the ACF+ model already reaches its boundaries.

D. Influence of the model size

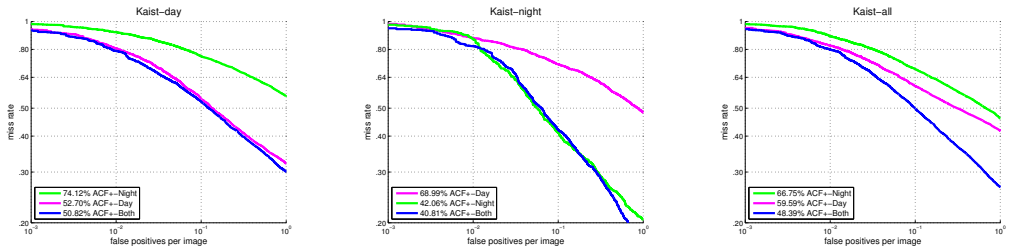
In [16] the authors encountered an accuracy benefit of using a larger model size when evaluating the ACF detector on the Caltech dataset. We have empirically tested if this also hold for our situation by training 5 additional models, going in steps of 10px from our 50px high model to a 100px high model. The accuracy obtained on the KAIST dataset is shown in figure 7. Although a slight accuracy improvement is obtained when using the 70px high model, we assume this improvement is rather a coincidence and not a better model setting in general, since it is not present with larger models. Also, the computation cost that comes with the necessity to upscale the source image when using a larger model, is not worth this small accuracy gain. Table 2 compares the detection speed when using different model sizes.

E. Using convolution masks

Based on the success of the use of convolution masks of [15], [16], [2], we extend our ACF+ model with the use of the 9 convolution masks of [2]. As figure 8 shows, do we get an accuracy gain using these masks, although not as large as we would expect. Since our training parameters differ from the ones used by the authors of [2], including the use of a larger feature pool, is it however possible this accuracy gain can be extended with other training parameters.

F. Using a ground constraint

As De Smedt et al. [19] demonstrated, it is possible to exploit a ground constraint by modelling the relation between the y-position in the image and the height of the pedestrian based on a training dataset, for which they used the Caltech dataset to validate this statement. Since the setting of this dataset is very similar to ours, we could also apply this technique here. Figure 8 visualises the influence of using a ground constraint on top of the improvements we gave earlier. We apply this ground constraint as a post-processing step, so we have no information on how large the speed improvement will be using this constraint here, but according to [19] a speed-up between 1.5 times and 3.9 times can be expected.



(a) test-set: KAIST-day (b) test-set: KAIST-night (c) test-set: KAIST-all

Fig. 4: Influence of trainingset vs testing set

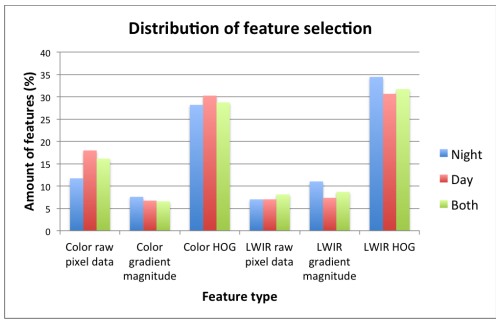


Fig. 5: Normalised feature distribution.

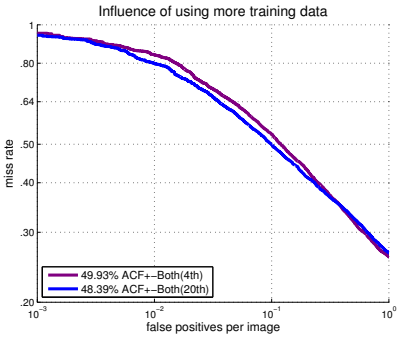


Fig. 6: Influence of the amount of training data

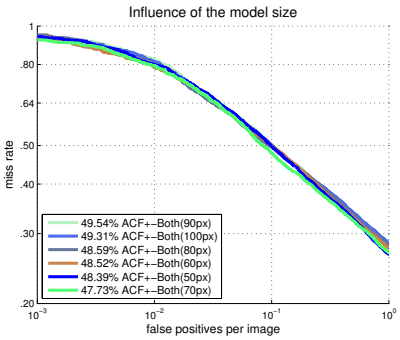


Fig. 7: Influence of the model size on detection accuracy.

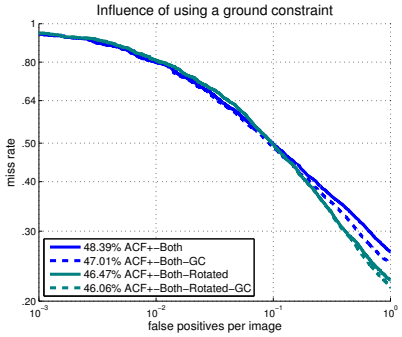


Fig. 8: Influence of using convolution masks and/or a ground constraint.

VI. System Test Experiments

In section III we described the use of a system test to validate a warning system for drivers when pedestrians are in front of the car. Here we apply this system test using the different models we used in the previous sections. Figure 9 shows an ROC-curve of the results. Note that instead of evaluating the pedestrian detection accuracy, we validate here if each frame is classified whether or not it should lead to an alarm. Table 2 gives a summary of the models used, with the associated detection speed using them. Based on the results in this table, we can see that the required speed of 1.85 fps under dry conditions can be reached, assuming a speed-up of the ground constraint. Under wet conditions however, the use of convolution masks have to undergo a large speed-up which is too severe for only using a ground constraint.

Note that the detector models are trained for full body detection, while the system test is independent to occlusion. When we evaluate only on night conditions, we get the remarkable result that we reach a TPR of 94% using the *ACF+-Both* model. Using only color information reaches very poor results, which we expected based on the results in figure 2. Although night conditions seem more complex on first sight, we have to remember that LWIR is especially useful in night conditions due to the high heat contrast between pedestrians and the surrounding temperature. The assistance of the driver, who is limited by using the visual spectrum, is especially useful in night conditions.

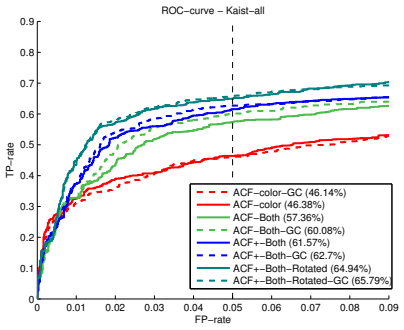


Fig. 9: Results of the system test on both day and night images.

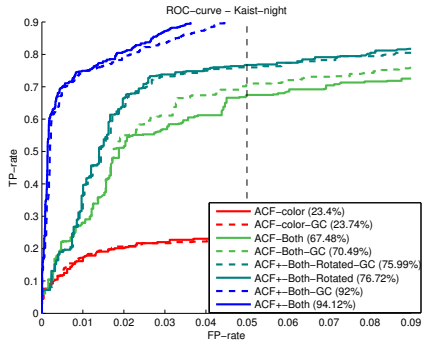


Fig. 10: Results of the system test on only night images.

VII. Conclusion

In this paper we proposed the use of current state-of-art pedestrian detection techniques as a warning system for car drivers. To validate this, we used a system test to validate if an alarm is generated when pedestrians are too close (less than 20m) in front of the car. Since the KAIST dataset, which is captured using a forward looking camera on a driving car, contains image pairs (color and LWIR) from both day and night conditions, it forms a suitable datataset for our experiments. We extended the ACF detector using only color information step by step by incorporating LWIR information and study the influence of different training choices. This way, we where capable of decreasing the miss-rate drastically compared to using only color, and outperform the current state-of-the-art on this dataset. According to our system tests on the combination of day and night conditions, we reached 65.79% TPR at a FPR of 5%,which is an almost 20% increase over using only color information. In night conditions the results of our system test reaches 94%. Especially at night conditions these results are very promising, since these are the conditions the driver would mostly benefit from a warning system.

Table 2: Speed comparison when using other model sizes

Method	$\geqslant$ 50px	$\geqslant$ 75px	$\geqslant$ 100px	TPR	TPR (GC)
ACF-color	10.73 fps	18.8 fps	26.03 fps	46.38	46.14
ACF-both	9.51 fps	11.81 fps	21.13 fps	57.36	60.08
ACF+ (50px)	8.75 fps	10.43 fps	19.28 fps	61.57	62.70
ACF+ (60px)	6.48 fps	7.58 fps	15.86 fps	60.08	62.73
ACF+ (70px)	5.06 fps	6.84 fps	13.48 fps	62.38	62.64
ACF+ (80px)	4.13 fps	8.14 fps	11.54 fps	61.76	61.11
ACF+ (90px)	3.33 fps	6.83 fps	10.29 fps	60.71	61.52
ACF+ (100px)	2.74 fps	5.56 fps	9.23 fps	59.66	59.43
ACF+ Rotated	0.875 fps	1.39 fps	1.91 fps	64.94	65.79

References

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005*, vol. 1, pp. 886–893, IEEE, 2005.

[2] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?," *arXiv preprint arXiv:1602.01237*, 2016.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, vol. 34, no. 4, pp. 743–761, 2012.

[4] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *CVPR 2015*, pp. 1037–1045, 2015.

[5] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR 2008*, pp. 1–8, IEEE, 2008.

[6] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *CVPR 2010*, pp. 2241–2248, IEEE, 2010.

[7] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *CVPR 2013*, pp. 3033–3040, 2013.

[8] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?," in *ECCV 2014 Workshops*, pp. 613–627, Springer, 2014.

[9] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.

[10] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[11] R. Benenson, M. Mathias, T. Tuytelaars, and L. Gool, "Seeking the strongest rigid detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3666–3673, 2013.

[12] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *PAMI*, vol. 36, no. 8, pp. 1532–1545, 2014.

[13] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west.," in *BMVC*, vol. 2, p. 7, Citeseer, 2010.

[14] R. Appel, T. Fuchs, P. Dollár, and P. Perona, "Quickly boosting decision trees-pruning underachieving features early," in *JMLR Workshop and Conference Proceedings*, vol. 28, pp. 594–602, JMLR, 2013.

[15] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, pp. 424–432, 2014.

[16] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *CVPR 2015*, pp. 1751–1760, IEEE, 2015.

[17] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *BMCV*, vol. 1, no. 3, p. 6, 2015.

[18] P. Sudowe and B. Leibe, "Efficient use of geometric constraints for sliding-window object detection in video," in *ICCV*, pp. 11–20, Springer, 2011.

[19] F. Smedt, D. Hulens, and T. Goedemé, "On-board real-time tracking of pedestrians on a uav," in *CVPR 2015 Workshops*, pp. 1–8, 2015.

[20] D. Hoedemaeker, M. Doumen, M. De Goede, J. Hogema, R. Brouwer, A. Wennemers, R. rapport Ongerubriceerd, S. Ongerubriceerd, R. Ongerubriceerd, B. Ongerubriceerd, *et al.*, *Modelopzet voor Dode-hoek Detectie en Signalerings Systemen (DDSS)*. Soesterberg: TNO; Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV), 2010.

[21] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *CVPR 2012*, pp. 2903–2910, IEEE, 2012.

[22] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *CVPR 2015*, pp. 4073–4082, 2015.

[23] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, 1998.